A Hierarchical Algorithm for Calculating the Isotopic Fine Structures of Molecules

Long Li,[a,b] Joshua A. Kresh,[b] N. Murat Karabacak,[a] Jennifer S. Cobb,[a] Jeffrey N. Agar,[a] and Pengyu Hong[b]

[a] Department of Chemistry and Volen Center for Complex Systems, Brandeis University, Waltham, Massachusetts, USA

[b] Department of Computer Science, National Center of Behavioral Genomics, Brandeis University, Waltham, Massachusetts, USA

This article presents a memory efficient algorithm for accurately calculating the isotopic fine structures of molecules. Treating individual isotopic species of a molecule as different mass states, we introduce the concept of transitions between mass states and represent all mass states of the molecule in a hierarchical structure. We show that there exists a simple relationship between two different mass states at two different levels of the hierarchical structure. This allows us to efficiently and accurately compute both the mass and the abundance of every mass state of a small to medium-sized molecule, whose gross structures include small number of fine structures. A truncated calculation of this algorithm can be applied to calculate a majority of isotopic species (99.99% of cumulative abundance) of a large molecule. (J Am Soc Mass Spectrom 2008, 19, 1867–1874) © 2008 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

To successfully interpret experimental mass spectrometry data, it is necessary to calculate the theoretical "continuous" isotopic envelope of a molecule for comparison. One strategy of the theoretical calculation is to first calculate the theoretical isotopic distribution, i.e., a set of the discrete isotopic species (each isotopic species has its own mass, abundance, and composition), and the theoretical isotopic envelope can be generated by convoluting the isotopic distributions with a peak shape function (e.g., Gaussian or Cauchy-Lorentz function) that accounts for instrument resolution. Another strategy is to inherently integrate the peak shape function to the calculation and directly generate the isotopic envelope. The former strategy is used in polynomial-based methods and the latter is used in Fourier transform-based methods (for more information of these methods, vide infra). Earlier mass spectrometry instruments could only deal with small molecules due to the difficulty of ionization of large molecules. It is easy to interpret the experimental spectra of a small molecule because its isotopic distribution is simple. The invention of electrospray ionization (ESI) [1] and matrix-assisted laser desorption/ionization (MALDI) [2] have made mass spectrometry of large biological molecules (e.g., proteins) possible [3, 4]. However, isotopic distribution/envelope becomes more complicated as the mass of the molecule increases. In fact, the abundance of the monoisotopic species of a large molecule (>100,000 Da) becomes vanishingly low, and is generally not observed. Moreover, a single isotopic species does not dominate in the isotopic distribution of such large molecules [5]. Furthermore, individual low abundance isotopic species contribute substantially to the isotopic distribution of large molecules. For instance, the most abundant isotopic species of bovine insulin is within the M + 2 nominal mass, however, the experimental highest peak is the M + 3 at low-resolution of $m/\Delta m_{50\%}$ 30,000 because there are more isotopic species within M + 3 than within M + 2 [6]. With the improvement of instrument resolution, the isotopic fine structures of large molecules can be observed. Not only the mass and the abundance, but the isotopic composition information is necessary to characterize the experimental spectra, such as the isotopic species assignment [7]. Moreover, it is theoretically important to discuss the calculation of the mass, abundance, and the isotopic composition of each isotopic species of a molecule.

The calculation of the isotopic distribution/envelope has been extensively studied since 1960. Starting from probability theory-based methods [8, 9] and mechanical methods [10, 11] for small molecules, it gradually evolved to polynomial expansion methods [12–14], which include a general fundamental principal for any molecule: the entire isotopic species of a molecule can be mathematically expressed in a concise and compact polynomial form:

Address reprint requests to either Dr. J. Agar, Department of Chemistry, MS 029, Brandeis University, 415 South Street, Waltham, MA 02454, USA, e-mail: agar@brandeis.edu; or to Dr. P. Hong, Department of Computer Science, MS 018, Brandeis University, 415 South Street, Waltham, MA 02454, USA. E-mail: hongpeng@brandeis.edu.

1044-0305/08/$32.00
doi:10.1016/j.jasms.2008.08.008

Published online August 15, 2008
Received April 17, 2008
Revised August 1, 2008
Accepted August 4, 2008

$$(p_{11}x^{m_{11}} + p_{12}x^{m_{12}} + \cdots)^{n_1} \cdots (p_{i1}x^{m_{i1}} + \cdots + p_{ij}x^{m_{ij}} + \cdots)^{n_i} \cdots \quad (1)$$

where m_{ij} and p_{ij} are the mass and the abundance of the *jth* isotope of the *ith* element, respectively; n_i is the number of all atoms of the *ith* element. Based on this polynomial representation, many stepwise calculation methods [15–20] have been proposed. All of these polynomial-based methods could theoretically compute the exact isotopic distribution of any molecule, i.e., "infinite" resolution, but fail for large molecules in practice because the number of isotopic species of a large molecule could be immense (due to combinatorial explosion with the increase of molecule weight). For example, polystyrene ($C_4H_9(C_8H_8)_{10,000}H$) and bovine insulin ($C_{254}H_{377}N_{65}O_{75}S_6$) have 6401,280,055 and 1563,613,904,160 isotopic species, respectively, if we only consider stable isotopes: ^{12}C and ^{13}C for carbon, ^{1}H and ^{2}H for hydrogen, ^{14}N and ^{15}N for nitrogen, ^{16}O, ^{17}O and ^{18}O for oxygen, ^{32}S, ^{33}S, ^{34}S and ^{36}S for sulfur. To calculate the exact mass and the exact abundance of every isotopic species of a molecule, polynomial-based methods need to keep all isotopic species of the molecule in memory simultaneously. Hence these methods require memory that is usually not available on modern personal computers. For example, more than 4000 gigabytes of RAM are required to compute the isotopic distribution of bovine insulin. To deal with the memory problem, stepwise methods used the pruning technique [15] each time an atom [20] or an element [15] or a hypothetical atom cluster [19] is added to only keep those isotopic species with abundances above a user defined threshold. Therefore, a direct consequence of pruning is that some species are missing. Moreover, possibly the deletion of a low abundance isotopic species at an early step could result in the deletion of one or more high abundance (greater than the threshold) isotopic species in subsequent steps. Although the masses and the abundances of remaining isotopic species are still exact, those missing isotopic species could result in significant distortion of the isotopic envelope especially for large molecules [21].

On the other hand, based on the convolution theorem, Rockwood and colleagues developed Fourier transform-based methods [21–24] to directly calculate the theoretic isotopic envelope. These Fourier transform-based methods are conceptually independent of polynomial-based methods, although they can be connected with each other [22]. One of the strengths of Fourier transform-based methods is that they do not suffer from the combinatorial explosion in the polynomial-based methods. Only a number of sampling data points (e.g., 2048) is needed to calculate the whole isotopic envelope at the resolution of $\sim 10^5$. When zooming into a limited mass range, using an array size of 2048 data points can achieve a resolution of 290,000,000 [24]. Therefore, Fourier transform-based methods are very efficient in memory. Moreover, the integrated abundance and the centroids of the calculated peaks are very close to theoretical value. In addition, by taking the advantage of fast Fourier transform algorithm, Fourier transform-based methods are computationally very fast. Because of these advantages, Fourier transform-based methods are superb at handling large molecules. For example, when calculating a DNA oligomer of molecular mass >123 kDa, the calculation completed in less than one second at a resolution of ~400,000 while an array size of only 4096 double precision points was used [21]. Compared to polynomial-based methods, the peak profile resulting from Fourier transform-based methods can be directly compared to the experimental isotopic envelope without requiring the additional computational time and effort of convoluting the peak width function to each isotopic species. One major drawback of Fourier transform-based methods is that they lose the isotopic composition information of each isotopic species. There is inherently a resolution parameter associated with the peak profile, which can result in a degradation of resolution of individual isotopic species. Rockwood et al. investigated the isotopic composition within each nominal mass peak [25]. The accurate mass and accurate abundance of the nominal mass could be calculated from the isotopic composition information. Still, the isotopic composition information of individual isotopic species is still missing.

There are another stepwise method [26] and an approximate method [27] to calculate the nominal peaks. Conversely, these methods do not provide the isotopic fine structure information of a molecule, and such information is required to fit experimental data from Fourier transform (see Figure 2) and sector instruments.

Inspired by the fine structure of atomic spectra and the transitions between different energy states in atomic physics, we developed a new method to accurately calculate the isotopic species of a molecule in a memory efficient way. We represent all isotopic species as different mass states, each of which is associated with a "configuration number" — its isotopic composition. The monoisotopic mass state is called the ground state, and the others are all excited states. Herein all mass states of a molecule can form a hierarchical structure that serves as a base for our simple recursive algorithm to calculate the entire isotopic distribution of the molecule.

Transition Theory and Algorithm

In atomic physics, a set of quantum numbers are associated with the energy states of the atom. The main electron shells of atoms, which are symbolized by the principal quantum number *n*, correspond to the gross structure of line spectra; the fine structures are caused by spin-orbit (which are symbolized by *s* and *l*, respectively) coupling and describe the splitting of the spectral lines of atoms.

The monoisotopic peak of a molecule has a unique elemental composition, i.e., all hydrogen are ^{1}H, all carbons are ^{12}C, etc. At approximately integer multiples

of ~1 Da higher in nominal mass, there are more than one isotopic composition which have the same nucleon number but differ by a few mDa. For example, at the nominal mass ~2 Da higher than the monoisotopic mass, the isotopic composition could be obtained by the change from two ^{12}C in monoisotopic composition to two ^{13}C, or two ^{14}N to two ^{15}N, or one ^{12}C to ^{13}C and one ^{14}N to one ^{15}N, and so on. By analogy, we use the nucleon number and the configuration number (isotopic composition) to define the isotopic gross structure and the isotopic fine structure of a molecule, respectively. All isotopic species of the molecule are considered as different mass states. Suppose for each mass state, its isotopic composition (which we call configuration number) is

$$[n_{11}, n_{12}, \cdots ; \cdots ; n_{i1}, n_{i2}, \cdots , n_{ij}, \cdots ; \cdots]$$

where n_{ij} stands for the number of the *jth* isotope for the *ith* element in the molecule. The isotopes (separated by the commas) of each element (separated by the semicolons) are sorted by their nucleon numbers, it means that n_{i1} is the lightest isotope of the *ith* element, n_{i2} is the second lightest isotope, and so on. The corresponding nucleon number, mass, and abundance of the mass state are *A*, *m*, and *p*, respectively:

$$A = A_{11}n_{11} + A_{12}n_{12} + \cdots + A_{i1}n_{i1} + A_{i2}n_{i2} + \cdots + A_{ij}n_{ij} + \cdots \quad (2)$$

$$m = n_{11}m_{11} + n_{12}m_{12} + \cdots + n_{i1}m_{i1} + n_{i2}m_{i2} + \cdots + n_{ij}m_{ij} + \cdots \quad (3)$$

$$p = \left(\frac{n_1!}{n_{11}!\,n_{12}!\cdots}p_{11}^{n_{11}}p_{12}^{n_{12}}\cdots\right)\cdots \left(\frac{n_i!}{n_{i1}!\cdots n_{ij}!\cdots}p_{i1}^{n_{i1}}\cdots p_{ij}^{n_{ij}}\cdots\right)\cdots \quad (4)$$

where $n_i = \sum_j n_{ij}$, i.e., the number of atoms of the *ith* element; A_{ij}, m_{ij}, and p_{ij} are the nucleon number, the mass, and the abundance of the *i*th isotope for the *j*th element, respectively.

We term each possible nucleon number A as a gross structure, which includes a set of fine structures, i.e., all mass states of the same nucleon number but different configuration numbers. Specifically, the mass state which consists of the lightest isotopes for all elements, i.e., $[n_1, 0, 0, \ldots ; n_2, 0, 0, \ldots ; \ldots]$ is referred to as the ground state, and the others are all excited states. Each molecule only has one ground state. Those mass states whose nucleon number is exactly one more than ground state are the first excited states, mass states of exactly two more nucleon numbers than ground state are the second excited states, and so on. Finally, there is only one highest excited-state, in which all of atoms are the heaviest isotopes for each element.

For example, carbon monoxide (CO) has four gross structures, i.e., 28, 29, 30, and 31, if we only consider the stable isotopes ^{12}C, ^{13}C, ^{16}O, ^{17}O, ^{18}O, and ignore unstable isotopes such as ^{14}C. The configuration number of ground state (nucleon number 28) is $^{12}C^{16}O/[1,0;1,0,0]$ (the digital numbers before the semicolon are the numbers of ^{12}C and ^{13}C, and those after the semicolon are the numbers of ^{16}O, ^{17}O and ^{18}O, respectively. The same symbols are used thereafter). There are two first-excited mass states (nucleon number 29), $^{12}C^{17}O/[1,0;0,1,0]$ and $^{13}C^{16}O/[0,1;1,0,0]$; there are two second-excited mass states (nucleon number 30), $^{12}C^{18}O/[1,0;0,0,1]$ and $^{13}C^{17}O/[0,1;0,1,0]$; for nucleon number 31, there is only one mass state, $^{13}C^{18}O/[0,1;0,0,1]$.

We define the element's "transition" as the conversion between two different isotopes of that element. If the element has *k* isotopes, the number of possible transitions is $k/(k-1)/2$. For example, carbon has only one transition: between ^{12}C and ^{13}C; oxygen has three transitions: between ^{16}O and ^{17}O, ^{17}O and ^{18}O, ^{16}O and ^{18}O.

We impose a "selection rule" on the transitions between the molecule's isotopic species: each time the transition is only allowed for just one atom of one element. For example, the mass state $^{12}C^{16}O/[1,0;1,0,0]$ can transmit to the following three mass states $^{12}C^{17}O/[1,0;0,1,0]$, $^{13}C^{16}O/[0,1;1,0,0]$ and $^{12}C^{18}O/[1,0;0,0,1]$, but not states $^{13}C^{17}O/[0,1;0,1,0]$ and $^{13}C^{18}O/[0,1;0,0,1]$, see Figure 1.

If we arrange all of mass states by nucleon number, they form a hierarchical structure (see Figure 1). The top level (level 0) only contains the lightest isotope (ground state) of that molecule; the bottom level only contains the heaviest isotope (highest excited-state) of that molecule. The nucleon number difference between any two adjacent levels is one. This hierarchical relationship, as shown later, allows the mass and the abundances of each mass state to be quickly computed in a memory efficient way.

Two variables are necessary to characterize both the isotopic gross structures and the isotopic fine struc-

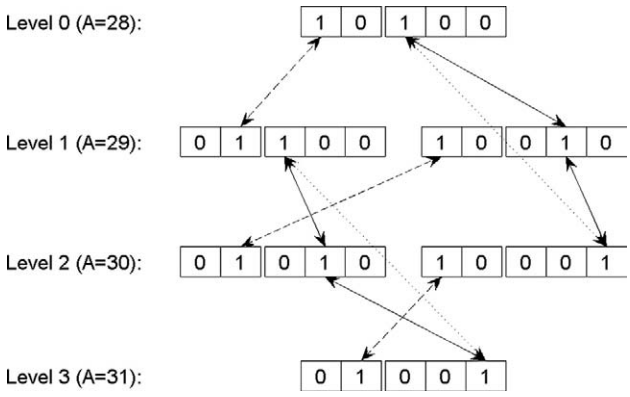

Figure 1. The transitions and hierarchical relationship between different mass states of carbon monoxide. The dashed lines stand for the conversion between ^{12}C and ^{13}C; the solid lines between ^{16}O and ^{17}O, or between ^{17}O and ^{18}O; dotted lines between ^{16}O and ^{18}O. Out of these transitions, only the conversion between ^{16}O and ^{18}O changes (dotted lines) traverse two levels, others only traverse one level.

tures. An isotopic fine structure denotes one individual mass state and is associated with its mass and abundance. An isotopic gross structure denotes one level and is associated with two statistics of the mass states within that level: (a) the average mass of all mass states weighted by their abundance and (b) the sum of the abundance of all mass states.

The hierarchical structure and the selection rule allow us to easily compute the mass and the abundance of each mass state. Let us start from level 0, which only includes the ground state for any molecule. The probability of the ground state can be simply written as

$$p_{11}^{n_1} p_{21}^{n_2} \cdots p_{i1}^{n_i} \cdots = \prod_i p_{i1}^{n_i} \tag{5}$$

where p_{i1} is the abundance of the lightest (or first) isotope of the *ith* element. In logarithm space, it becomes:

$$n_1 \log_{10} p_{11} + n_2 \log_{10} p_{21} + \cdots + n_i \log_{10} p_{i1} + \cdots = \sum_i n_i \log_{10} p_{i1} \tag{6}$$

The mass of ground state can be simplified to:

$$n_1 m_{11} + n_2 m_{21} + \cdots + n_i m_{i1} + \cdots = \sum_i n_i m_{i1} \tag{7}$$

Because of the selection rule, the composition of the two mass states of occurring transition change very little, i.e., from $[\cdots;\cdots,n_{ij},\cdots,n_{ij'},\cdots;\cdots]$ to $[\cdots;\cdots,n_{ij}-1,\cdots,n_{ij'}+1,\cdots;\cdots]$. Starting from level 1, the following recursive formulas are used to calculate the mass and probability:

$$m([\cdots;\cdots,n_{ij}-1,\cdots,n_{ij'}+1,\cdots;\cdots]) = m([\cdots;\cdots,n_{ij},\cdots,n_{ij'},\cdots;\cdots]) + m_{ij'} - m_{ij} \tag{8}$$

$$p([\cdots;\cdots,n_{ij}-1,\cdots,n_{ij'}+1,\cdots;\cdots]) = p([\cdots;\cdots,n_{ij},\cdots,n_{ij'},\cdots;\cdots])\frac{n_{ij}p_{ij'}}{n_{ij'}p_{ij}} \tag{9}$$

In logarithm space, the probability formula becomes

$$\log_{10} p([\cdots;\cdots,n_{ij}-1,\cdots,n_{ij'}+1,\cdots;\cdots]) = \log_{10} p([\cdots;\cdots,n_{ij},\cdots;n_{ij'}\cdots;\cdots]) + \log_{10} n_{ij} + \log_{10} p_{ij'} - \log_{10} n_{ij'} - \log_{10} p_{ij} \tag{10}$$

These recursive formulas are similar to Yergey's method [15], where they are applied to a single element. They allow us to avoid repeatedly calculating the factorial (the factorial evaluations often cause overflow for large whole numbers) and the exponential part in eq 4 for each state.

Discrete isotopic species could not compare with experimental spectra directly. We generate the theoretical isotopic envelope for comparison by convoluting a peak shape function to the discrete isotopic species. The most often used peak shape functions are the Gaussian function and the Cauchy-Lorentz function. The contribution of each isotopic species (it mass is m_k and abundance is p_k) to the whole envelope can be calculated by:

$$f(m;\sigma) = p_k \exp\left(-\frac{(m-m_k)^2}{2\sigma^2}\right) \tag{11}$$

$$f(m;\gamma) = \frac{\gamma^2 p_k}{\gamma^2 + (m-m_k)^2} \tag{12}$$

Where σ and γ are the parameters of the Gaussian and Cauchy-Lorentz functions, respectively, which are used to control the width of the peak. Their full width at half maximum are $\sigma(\ln 256)^{1/2}$ and 2γ, respectively. Therefore, their resolutions defined by $R = m/\Delta m_{50\%}$ are $m_k/(\sigma(\ln 256)^{1/2})$ and $m_k/2\gamma$, respectively. The final form of the formula to calculate the theoretical isotopic envelope is:

$$f_{total}(m) = N\sum_k p_k \exp\left(-\frac{(m-m_k)^2 R^2 \ln 256}{2m_k^2}\right) \tag{13}$$

$$f_{total}(m) = N\sum_k \frac{p_k m_k^2}{m_k^2 + 4R^2(m-m_k)^2} \tag{14}$$

where N is the normalization to experimental spectra. In practice, to generate a experiment-comparable envelope, we take a number of equal-interval points within a chosen mass range (e.g., the left bound is 1 Da less than the mass of the lightest isotope species and the right bound is 1 Da more than the mass of the highest isotope species; the masses of these two species are easily calculated), then use equations above to calculate the corresponding abundance at each point. Note, here these sampling data points have nothing to do with the resolution. In the Fourier transform-based method, the sampling data points are related to the resolution: the more the sampling data points, the higher the resolution.

Results and Discussion

We applied our new algorithm to human neuropeptide substance P with two additional protons ($C_{63}H_{100}N_{18}O_{13}S$) and the theoretical calculated isotopic envelope at the resolution of $m/\Delta m_{50\%}$ 800,000 agrees with the experimental data well; see Figure 2. $C_{63}H_{100}N_{18}O_{13}S$ has 51,582,720 individual isotopic species spread across 212 levels, and it only takes about 40 min for our algorithm to generate the whole isotopic distribution on Intel Core2 Duo CPU@2.0GHz machine with 2 gigabytes of RAM. As a comparison, the latest published isoDalton program [20] ran out of memory on the same machine (isoDalton was downloaded from MATLAB Central File Exchange that

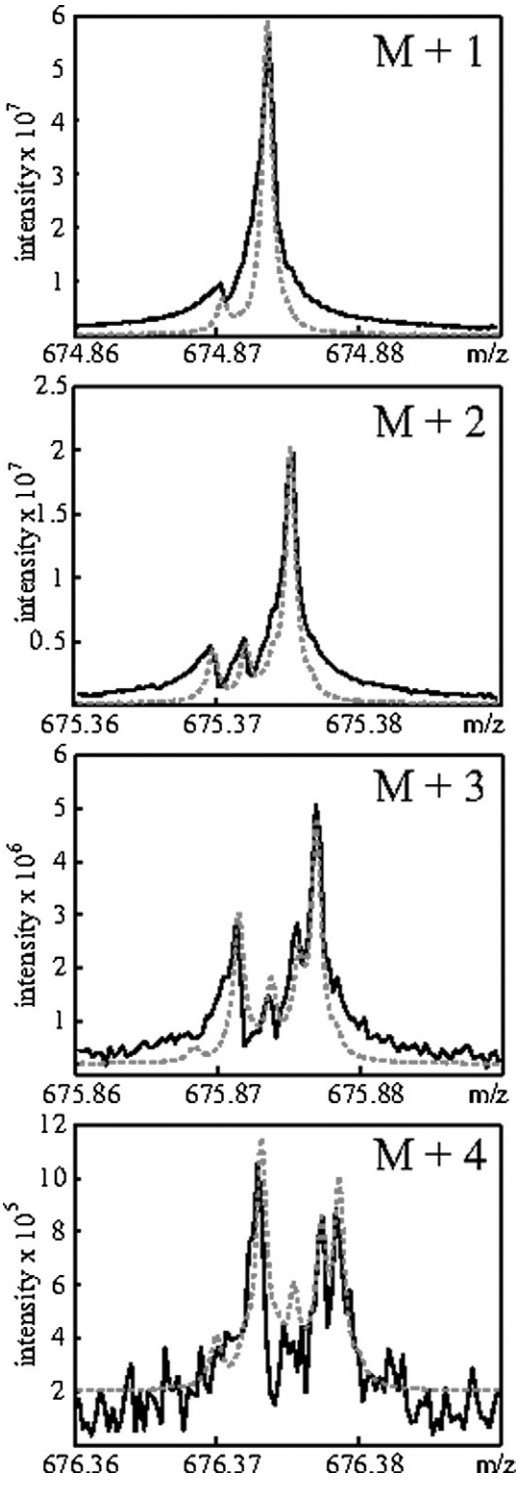

Figure 2. The theoretical calculated isotopic envelope (dotted line) generated from our algorithm (with Cauchy-Lorentz peak function) and the Fourier transform-ion cyclotron resonance mass spectrum (solid line) of human neuropeptide substance P with two additional protons. Only M + 1, M + 2, M + 3, and M + 4 are shown; M has no fine structure. The resolution for the theoretical calculated envelope is $m/\Delta m_{50\%}$ 800,000.

was submitted on 07/30/2007, http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId = 15,752).

For any transition in our algorithm, only the mass difference between two different isotopes of the same element is added to get the new mass, and the same is applied to the logarithm of the abundance. We can easily compute the abundance and the mass of the ground state using eqs 6 and 7. The remaining mass states can be iteratively computed from the ground state using eqs 9 and 10. For example, the first excited states in level one can be calculated based on the ground state; the second excited states in level two can be calculated based on the first excited states, and so on.

Moreover, compared with previous polynomial-based methods, our method improves the memory efficiency by dispersing the whole mass state set to different levels when calculating the discrete isotopic species. Most of the time we only need to keep two adjacent levels in the memory during the whole calculation process (for those elements whose isotopes are not continuous, for example, sulfur has ^{32}S, ^{33}S, ^{34}S, and ^{36}S, but no stable ^{35}S, three levels have to be kept in the memory to compensate the jump from ^{34}S to ^{36}S). As a result, the number of mass states stored in the memory is much smaller than the size of the whole mass state set. Note that the number of states at each level starts with one at the level including only ground state, it first increases with the depth of the level, then decreases, and eventually becomes one at the bottom level containing only the highest excited-state. Therefore, the isotopic distributions of large molecules can be accurately computed on a typical personal computer. For instance, molecule $C_{100,000}H_{200,000}$ (1,301,870 Da) has 10,000,200,001 different mass states, which are spread across 200,001 levels. Previous polynomial-based meth-

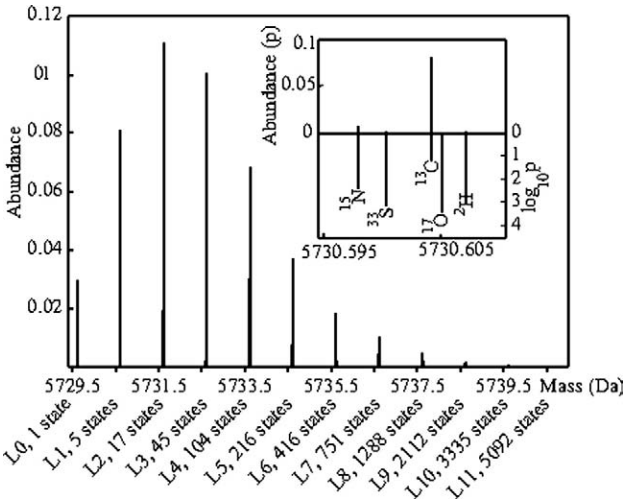

Figure 3. The calculated isotopic distribution in the first 12 main levels (including 13,382 mass states) for bovine insulin (totally 1563,613,904,160 mass states dispersed into 871 levels), these 12 main levels have already represented 99.99% abundance of the whole isotopic species. The inset panel shows all of five states in Level 1: the upper part is the absolute abundance for each state and the lower part is the logarithm of the absolute abundance; the isotopic change for each state relative to the ground state is also shown.

ods will run out of memory to calculate the whole isotopic distribution. In our method, the maximum number of the mass states in a level is only 100,001. Hence, our method requires only several megabytes of memory for this molecule.

Although our algorithm significantly saves memory, it does not completely solve the memory problem because the number of the mass states at a single level could still be too large to be kept in the memory. For example, the maximum number of the mass states in a single level of bovine insulin is 4.1×10^9. A machine with 2 gigabytes of RAM could not compute the whole isotopic distribution. Hence we also offer a truncation option to calculate the major mass states of molecules. It is based on the observation that the first few levels in the hierarchical structure include most high abundance isotopic species of a molecule because the isotopic abundances of main group elements are heavily biased to the lightest isotopes, e.g., about 98.93% carbon atoms are ^{12}C. Therefore, there exists a level G so that the abundance of any mass state beyond level G is very close to zero, and at the same time the sum of the abundances of the mass states beyond level G is also close to zero. We can terminate the computation at this level and truncate the other levels. For example, in the hierarchical structure of bovine insulin there are in total 871 levels and 1563,613,904,160 mass states. However, the mass states in the first 12 levels (including 13,382 mass states) have already represented 99.99% of the cumulative probability distribution (see Figure 3). As mentioned in the introduction, previous pruning techniques discard the isotopic species with low abundances at each level (sometimes even the isotopic species with high abundances are discarded), which results in distortion of the whole isotopic envelope. In our algorithm, every isotopic species in each level before truncation is retained; therefore we can avoid the distortion problem. This truncation reduces a significant portion of the calculation time. For instance, it takes less than one second to calculate the first 12 levels of bovine insulin. The truncation level G can be determined by the cumulative abundance from level zero to level G. Our experience is that cumulative abundance 99.9% is enough to generate a high quality isotopic envelope. In addition, the truncated calculation also saves the storage space. Although only the mass states in two or three levels are needed to be kept in memory, all computed mass states need to be output onto hard disk. The reasons for output onto hard disk will be explained as below. The more the mass states of a molecule, the larger the storage space needed. For example, saving all mass states of bovine insulin in binary format needs more than 8 terabytes space (ASCII format need more), but saving the mass states in first 12 levels only need ~150 kilobytes space. Usually the number of the mass states in the single level is small if the elements in the molecule have one or two isotopes, such as carbon or hydrogen or nitrogen. The truncation method can deal with this kind of "simple" molecules to several hundreds of kilodaltons. If the molecule consists of multiple-isotope elements like oxygen and sulfur, the number of the mass states in each level can rapidly increase with the molecular weight. We calculated different proteins in the IPI Human database [28] (ver. 3.31) and found, conservatively speaking, that the truncation method can deal with proteins up to 20 kDa on a personal computer.

The whole isotopic envelope could be generated on-the-fly: each time we obtain a new isotopic species, eq 11 or eq 12 is used to calculate the contribution of this species. In practice, we use the following two-step procedure to calculate the isotopic envelope. First we output the isotopic species to a file on the disk, then sort them by the mass (sometimes it is mandatory to save the mass, abundance, and composition information of individual isotopic species on the hard disk). For example, such information is used in the assignment of fine structure [7]. Sorting can be done quickly (usually less than one second) because those isotopic species have been semi-sorted by the level during the calculation. At the second step, the envelope is generated on-the-fly by loading the sorted isotopic species one by one, which saves a lot of memory because only the sampling data points are kept in the memory. This two-step procedure improves the speed as well. It will take longer time to compute all sampling points in the mass range when using eq 11 or eq 12 to calculate the contribution of single isotopic species. However, given that the peak function has the limited width at a certain resolution, only a portion of sampling points (e.g., the points within $\pm 5\sigma$ of central mass if we use Gaussian function) need to be calculated. Moreover, $\pm 5\sigma$ boundary for each isotopic species could be anchored efficiently because all of isotopic species have already been sorted. This is especially efficient for high-resolution cases because the width of peak function is small. For instance, it takes 68.17 s to calculate the isotopic envelope of $C_{10,000}H_{10,000}$ at the resolution of 10^4, while 7.3 s at the resolution of 10^7 for the same sample points (mass range: [130078 Da, 130213 Da], interval 0.0001 Da).

We compared our method with two other programs. One is the IsoPro [29], which implemented Yergey's algorithm [15], another is Mercury (kindly provided by Professor Rockwood), which incorporated a high-resolution profile-mode Fourier transform algorithm [21] and an ultrahigh resolution Fourier transform algorithm [24]. Although IsoPro does not provide exact calculation time, in our experience it is fast to deal with small and medium (<10,000 Da) peptides/proteins. For example, both the calculation of peak list and the generation of envelope of bovine insulin are done within 1 s. We did not use IsoPro to calculate large molecules because pruning used by polynomial-based methods including IsoPro can result in severe distortion, which has been discussed in [21]. The calculation time and memory usage for calculating a medium molecule $C_{254}H_{377}N_{65}O_{75}S_6$ and a large molecule $C_{10,000}H_{10,000}$ using Mercury with default parameters and our truncated cal-

Table 1. The comparison between Fourier transform-based methods and our method in computation time and memory usage

	Truncated calculation		Mercury	
	$R = 10^5$	$R = 10^7$	High resolution mode ($R:10^4 \sim 10^5$)	Ultrahigh resolution mode ($R > 10^7$)
$C_{254}H_{377}N_{65}O_{75}S_6$				
Time	3.26 s	0.78 s	<0.05 s	$7.11\ s\left(X\frac{5739.8-5729.5}{0.2}=51.5\right)$
Memory	~0.5 Mb		~5 kb	
$C_{10000}H_{10000}$				
Time	68.4 s	7.52 s	<0.05 s	$15.73\ s\left(X\frac{130212-130078}{0.2}=665\right)$
Memory	~1 Mb		~5 kb	

In our method, we use 99.9% cumulative abundance as the threshold to calculate the isotopic species, and the envelopes are generated on the mass ranges [5729.5, 5739.8] for $C_{254}H_{377}N_{65}O_{75}S_6$ and [130078, 130213] for $C_{10000}H_{10000}$, respectively, and interval 0.0001 Da. The calculation of Mercury was done using default parameters. At the ultrahigh resolution mode of Mercury, the calculation was done only within a default mass range of 0.2 Da (7.11s for $C_{254}H_{377}N_{65}O_{75}S_6$) near the average mass, and the total time for calculating the same mass range to our truncated method should multiply by the factor in the parenthesis (for example, the mass range of $C_{254}H_{377}N_{65}O_{75}S_6$ is 10.3 Da, therefore the total calculation time is 7.11s × 51.5).

culation, respectively, are listed in the Table 1. At the resolution of $\sim 10^5$, Mercury is too fast to record a reliable length of time. However, at the resolution $>10^7$, our method needs less time (Mercury code needs to be modified to use a larger array to get higher resolution. For example, the resolution is ~60,000 when calculating bovine insulin using the current array of 2048. The resolution can be reached $\sim 10^8$ if using an array of size ~16 M. This, however, will increase the calculation time (personal communications with Professor Rockwood). The calculation time in the second step of our algorithm (calculating the envelope) will decrease dramatically with the increase of resolution because high-resolution requires fewer points around each isotopic species. Mercury is more efficient in memory because it only requires a number of sampling data points (2048 in current version, i.e., ~5 kilobytes memory) is needed. In the second envelope generation step of our method, it can reach the same efficiency by using the same number of sampling data points to Mercury, while in the truncated calculation of isotopic species, usually ~1 megabytes memory is used.

Our current implement can calculate the theoretical isotopic distribution of protein molecules, which include the elements of C, H, O, N, and S. The program was written in C++ and tested on Linux. It is freely available under the GNU Lesser General Public License (http://www.cs.brandeis.edu/~hong/software.htm).

Conclusions

We developed a memory efficient algorithm for accurately calculating the entire isotopic fine structures of large molecules, on which previous methods run out of memory. The truncation option is offered to calculate the major isotopic fine structures and reduce the calculation time.

Acknowledgments

The authors are grateful to Professor Rockwood for kindly providing the Mercury program and the in-depth instructions. They thank Dr. N. Agar, Dr. M. Ravi Kumar, K. Boggio, J. Johnson, and Q. Wang for useful discussion and the comments on the manuscript. This work was supported by Brandeis Faculty Fund to P.H., in part by an award from the DOD (contract W81XWH-04-0158), and grant 1392 from the Amyotrophic Lateral Sclerosis Society of America to J.N.A.

References

1. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. *Science* **1989,** *246,* 64–71.
2. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60,* 2299–2301.
3. Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003,** *422,* 198–207.
4. Cravatt, B. F.; Simon, G. M.; Yates, J. R. The Biological Impact of Mass-Spectrometry-Based Proteomics. *Nature* **2007,** *450,* 991–1000.
5. Yergey, J.; Heller, D.; Hansen, G.; Cotter, R.; Fenselau, C. Isotopic Distributions in Mass Spectra of Large Molecules. *Anal. Chem.* **1983,** *55,* 353–356.
6. Raymond, C. W. Effect of Resolution on the Shape of Mass Spectra of Proteins: Some Theoretical Considerations. *Rapid Commun. Mass Spectrom.* **1994,** *8,* 976–980.
7. Shi, S. D.; Hendrickson, C. L.; Marshall, A. G. Counting Individual Sulfur Atoms in a Protein by Ultrahigh-Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: Experimental Resolution of Isotopic Fine Structure in Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95,* 11532–11537.
8. Beynon, J. H. *Mass Spectrometry and Its Applications to Organic Chemistry;* Elsevier: Amsterdam, 1960, 290–302.
9. Margrave, J. L.; Polansky, R. B. Relative Abundance Calculations for Isotopic Molecular Species. *J. Chem. Edu.* **1962,** *39,* 335–337.
10. Carrick, A.; Glocklin. F. Mass and Abundance Data for Polyisotopic Elements. J. Chem. Soc. A Inorg. Phys. Theor. **1967,** 40–42.
11. Robinson, R. J.; Warner, C. G.; Gohlke, R. S. Calculation of Relative Abundance of Isotope Clusters in Mass Spectrometry. *J. Chem. Educ.* **1970,** *47,* 467–468.
12. Hugentob, E.; Loliger J. General Approach to Calculating Isotope Abundance Ratios in Mass Spectroscopy. *J. Chem. Educ.* **1972,** *49,* 610–612.
13. Brownawell, M. L.; Filippo, J. S. A Program for the Synthesis of Mass-Spectral Isotopic Abundances. *J. Chem. Educ.* **1982,** *59,* 663–665.
14. Olsen, C. E. A Pascal Program for Micro-Computers for Calculations of Compositions and Isotope Clusters from Accurate Mass Measurements. *Int. J. Mass Spectrom Ion Processes* **1983,** *47,* 337–340.
15. Yergey, J. A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys.* **1983,** *52,* 337–349.

16. Hsu, C. S. Diophantine Approach to Isotopic Abundance Calculations. *Anal. Chem.* **1984,** *56*, 1356–1361.
17. Hibbert, D. B. A Prolog Program for the Calculation of Isotope Distributions in Mass-Spectrometry. *Chem. Intelligent Lab. Syst.* **1989,** *6*, 203–212.
18. Datta, B. P. Polynomial Method of Molecular Isotopic Abundance Calculations: A Computational Note. *Rapid Commun. Mass Spectrom.* **1997,** *11*, 1767–1774.
19. Kubinyi, H. Calculation of Isotope Distributions in Mass Spectrometry. A Trivial Solution for a Nontrivial Problem. *Anal. Chim. Acta* **1991,** *247*, 107–119.
20. Snider, R. K. Efficient Calculation of Exact Mass Isotopic Distributions. *J. Am. Soc. Mass Spectrom.* **2007,** *18*, 1511–1515.
21. Rockwood, A. L.; Vanorden, S. L.; Smith, R. D. Rapid Calculation of Isotope Distributions. *Anal. Chem.* **1995,** *67*, 2699–2704.
22. Rockwood, A. L. Relationship of Fourier-Transforms to Isotope Distribution Calculations. *Rapid Commun. Mass Spectrom.* **1995,** *9*, 103–105.
23. Rockwood, A. L.; VanOrden, S. L. Ultrahigh-Speed Calculation of Isotope Distributions. *Anal. Chem.* **1996,** *68*, 2027–2030.
24. Rockwood, A. L.; Van Orden, S. L.; Smith, R. D. Ultrahigh Resolution Isotope Distribution Calculations. *Rapid Commun. Mass Spectrom.* **1996,** *10*, 54–59.
25. Rockwood, A. L.; Van Orman, J. R.; Dearden, D. V. Isotopic Compositions and Accurate Masses of Single Isotopic Peaks. *J. Am. Soc. Mass Spectrom.* **2004,** *15*, 12–21.
26. Rockwood, A. L.; Haimi, P. Efficient Calculation of Accurate Masses of Isotopic Peaks. *J. Am. Soc. Mass Spectrom.* **2006,** *17*, 415–419.
27. Wang, B.; Sun, G.; Anderson, D. R.; Jia, M.; Previs, S.; Anderson, V. E. Isotopologue distributions of peptide product ions by tandem mass spectrometry: Quantitation of low levels of deuterium incorporation. *Anal. Biochem.* **2007,** *367*, 40–48.
28. Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: An Integrated Database for Proteomics Experiments. *Proteomics* **2004,** *4*, 1985–1988.
29. IsoPro, 3.0, http://members.aol.com/msmssoft/.